Solution 1: Risk Minimizers for 0-1-Loss

(a) The empirical risk of any $h \in \mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y} \,|\, h(\mathbf{x}) = \boldsymbol{\theta} \; \forall \mathbf{x} \in \mathcal{X}\}$ for the 0-1-loss, i.e.,

$$L\left(y, h(\mathbf{x})\right) = \mathbb{1}_{\{y \neq h(\mathbf{x})\}} = \begin{cases} 1, & \text{if } y \neq h(\mathbf{x}), \\ 0, & \text{if } y = h(\mathbf{x}), \end{cases}$$

is

$$\begin{aligned} \mathcal{R}_{\text{emp}}(h) &= \sum_{i=1}^{n} \mathbb{1}_{\{y^{(i)} \neq h\left(\mathbf{x}^{(i)}\right)\}} \\ &= \sum_{i:y^{(i)}=1} \mathbb{1}_{\{1 \neq h\left(\mathbf{x}^{(i)}\right)\}} + \sum_{i:y^{(i)}=2} \mathbb{1}_{\{2 \neq h\left(\mathbf{x}^{(i)}\right)\}} + \ldots + \sum_{i:y^{(i)}=g} \mathbb{1}_{\{g \neq h\left(\mathbf{x}^{(i)}\right)\}} \\ &= \sum_{j=1}^{g} \sum_{i:y^{(i)}=j} \mathbb{1}_{\{j \neq h\left(\mathbf{x}^{(i)}\right)\}} \\ &= \sum_{j=1}^{g} \sum_{i:y^{(i)}=j} \mathbb{1}_{\{j \neq \boldsymbol{\theta}\}}. && \text{(constant model)} \end{aligned}$$

For any $\boldsymbol{\theta} \in \mathcal{Y}$ and $j \in \mathcal{Y}$ we write

$$m_j(\boldsymbol{\theta}) = \sum_{i:y^{(i)}=j} \mathbb{1}_{\{j \neq \boldsymbol{\theta}\}},$$

i.e., the number of mistakes over the data set for class j by predicting $\boldsymbol{\theta}$. Further let

$$n_j = \sum_{i=1}^{n} \mathbb{1}_{\{y^{(i)}=j\}}$$

be the number of occurrences of the class j in the data set. If $k = \boldsymbol{\theta}$, then $m_k(\boldsymbol{\theta}) = 0$ and for any $j \neq k$ it holds that $m_j(\boldsymbol{\theta}) = n_j$. In words, if $\boldsymbol{\theta}$ coincides with k, then we make no mistake for this class k, while *for all* other classes j we make each time a mistake.

Let j^* be the mode of $y^{(1)}, \ldots, y^{(n)}$, i.e., the class which appears the most[1]. Note that by definition $n_{j^*} \geq n_j$ for any $j \neq j^*$. With this, we obtain that

$$\mathcal{R}_{\text{emp}}(h) = \sum_{j=1}^{g} \sum_{i:y^{(i)}=j} \mathbb{1}_{\{j \neq \boldsymbol{\theta}\}} = \sum_{j \neq \boldsymbol{\theta}} n_j \geq \sum_{j \neq j^*} n_j = \mathcal{R}_{\text{emp}}(\hat{h}),$$

since $\hat{h}(\mathbf{x}) = \text{mode}\left\{y^{(i)}\right\} = j^*$.

(b) Recall that the point-wise optimizer for the 0-1-loss over all possible discrete classifiers $h(\mathbf{x})$ is

$$h^*(\mathbf{x}) \quad = \quad \arg\max_{l \in \mathcal{Y}} \mathbb{P}(y = l \mid \mathbf{x} = \mathbf{x}).$$

Hence, we obtain the optimal constant model from the previous by forgetting the conditioning on $\mathbf{x}$, which leads to

$$\bar{h}(\mathbf{x}) = \arg\max_{l \in \mathcal{Y}} \mathbb{P}(y = l).$$

Recall that we can write the 0-1-loss as follows:

$$L\left(y, h(\mathbf{x})\right) = \mathbb{1}_{\{y \neq h(\mathbf{x})\}} = \sum_{k \in \mathcal{Y}} \mathbb{1}_{\{y=k\}} \mathbb{1}_{\{y \neq h(\mathbf{x})\}} = \sum_{k \in \mathcal{Y}} \mathbb{1}_{\{y=k\}} \mathbb{1}_{\{k \neq h(\mathbf{x})\}} = \sum_{k \in \mathcal{Y}} \mathbb{1}_{\{y=k\}} L(k, h(\mathbf{x})). \tag{1}$$

[1] Break ties arbitrarily.

With this, the risk of $\bar{h}$ is

$$\begin{aligned}
\mathcal{R}_L(\bar{h}) &= \mathbb{E}_{xy}\left[L(y,\bar{h}(\mathbf{x}))\right] \\
&= \mathbb{E}_x\left[\mathbb{E}_{y|x}[L(y,\bar{h}(\mathbf{x})) \mid \mathbf{x}=\mathbf{x}]\right] && \text{(Law of total expectation)} \\
&= \mathbb{E}_x\left[\mathbb{E}_{y|x}\left[\sum_{k\in\mathcal{Y}} \mathbb{1}_{\{y=k\}} L(k,\bar{h}(\mathbf{x})) \mid \mathbf{x}=\mathbf{x}\right]\right] && \text{(By (1))} \\
&= \mathbb{E}_x\left[\sum_{k\in\mathcal{Y}} L(k,\bar{h}(\mathbf{x}))\mathbb{E}_{y|x}\left[\mathbb{1}_{\{y=k\}} \mid \mathbf{x}=\mathbf{x}\right]\right] && \text{(Linearity of cond. expectation)} \\
&= \mathbb{E}_x\left[\sum_{k\in\mathcal{Y}} L(k,\bar{h}(\mathbf{x}))\mathbb{P}(y=k \mid \mathbf{x}=\mathbf{x})\right] && \text{(Expectation of an indicator random variable)} \\
&= \sum_{k\in\mathcal{Y}} \mathbb{E}_x\left[L(k,\bar{h}(\mathbf{x}))\mathbb{P}(y=k \mid \mathbf{x}=\mathbf{x})\right] && \text{(Linearity of expectation)} \\
&= \sum_{k\in\mathcal{Y}} L(k,\bar{h}(\mathbf{x}))\mathbb{E}_x\left[\mathbb{P}(y=k \mid \mathbf{x}=\mathbf{x})\right] && \text{($\bar{h}$ is constant in $\mathbf{x}$)} \\
&= \sum_{k\in\mathcal{Y}} L(k,\bar{h}(\mathbf{x}))\mathbb{P}(y=k) && \text{(Law of total probability)} \\
&= \sum_{k\in\mathcal{Y}} \mathbb{1}_{\{k\neq h(\mathbf{x})\}}\mathbb{P}(y=k) \\
&= \sum_{k\in\mathcal{Y}} \mathbb{1}_{\{k\neq \arg\max_{l\in\mathcal{Y}} \mathbb{P}(y=l)\}}\mathbb{P}(y=k) \\
&= 1-\max_{l\in\mathcal{Y}} \mathbb{P}(y=l).
\end{aligned}$$

(c) By recalling the definition of the approximation error:

$$\begin{aligned}
\inf_{h\in\mathcal{H}} \mathcal{R}_L(h) - \mathcal{R}_L^* = \quad &\underbrace{\mathcal{R}_L(\bar{h})}_{\overset{(b)}{=} 1-\max_{l\in\mathcal{Y}} \mathbb{P}(y=l)} \quad - \quad \underbrace{\mathcal{R}_L^*}_{\overset{Lec.}{=} 1-\mathbb{E}_x[\max_{l\in\mathcal{Y}} \mathbb{P}(y=l \mid \mathbf{x}=\mathbf{x})]} \\
&= \mathbb{E}_x\left[\max_{l\in\mathcal{Y}} \mathbb{P}(y=l \mid \mathbf{x}=\mathbf{x})\right] - \max_{l\in\mathcal{Y}} \mathbb{P}(y=l).
\end{aligned}$$

(d) For any probabilistic classifier π in the hypothesis space of probabilistic classifiers $\mathcal{H}=\{\pi:\mathcal{X}\to[0,1]\}$, we can write the probabilistic 0-1-loss as

$$L\left(y,\pi(\mathbf{x})\right) = \mathbb{1}_{\{\pi(\mathbf{x})\geq 1/2\}}\mathbb{1}_{\{y=0\}} + \mathbb{1}_{\{\pi(\mathbf{x})<1/2\}}\mathbb{1}_{\{y=1\}} \left(= \begin{cases} 1, & \text{if } (\pi(\mathbf{x})\geq 1/2 \text{ \& } y=0) \text{ or } (\pi(\mathbf{x})<1/2 \text{ \& } y=1), \\ 0, & \text{else.} \end{cases} \right).$$

We use our usual “unraveling trick” by means of the law of total expectation:

$$\mathbb{E}_{xy}\left[L(y,\pi(\mathbf{x}))\right] = \mathbb{E}_x\left[\mathbb{E}_{y|x}[L(y,\pi(\mathbf{x})) \mid \mathbf{x}=\mathbf{x}]\right]$$

and consider then minimization of $\mathbb{E}_{y|x}[L(y,\pi(\mathbf{x})) \mid \mathbf{x}=\mathbf{x}]$ by choosing π point-wise, i.e., for any point $\mathbf{x}$. With the alternative form of L, we obtain

$$\begin{aligned}
&\mathbb{E}_{y|x}[L(y,\pi(\mathbf{x})) \mid \mathbf{x}=\mathbf{x}] \\
&\quad= \mathbb{E}_{y|x}[\mathbb{1}_{\{\pi(\mathbf{x})\geq 1/2\}}\mathbb{1}_{\{y=0\}} + \mathbb{1}_{\{\pi(\mathbf{x})<1/2\}}\mathbb{1}_{\{y=1\}} \mid \mathbf{x}=\mathbf{x}] \\
&\quad= \mathbb{E}_{y|x}[\mathbb{1}_{\{\pi(\mathbf{x})\geq 1/2\}}\mathbb{1}_{\{y=0\}} \mid \mathbf{x}=\mathbf{x}] \quad + \mathbb{E}_{y|x}[\mathbb{1}_{\{\pi(\mathbf{x})<1/2\}}\mathbb{1}_{\{y=1\}} \mid \mathbf{x}=\mathbf{x}] \quad \text{(Linearity of expectation)} \\
&\quad= \mathbb{1}_{\{\pi(\mathbf{x})\geq 1/2\}} \cdot \mathbb{E}_{y|x}[\mathbb{1}_{\{y=0\}} \mid \mathbf{x}=\mathbf{x}] \quad + \mathbb{1}_{\{\pi(\mathbf{x})<1/2\}} \cdot \mathbb{E}_{y|x}[\mathbb{1}_{\{y=1\}} \mid \mathbf{x}=\mathbf{x}] \\
&\qquad\qquad \text{($\mathbb{1}_{\{\pi(\mathbf{x})\geq 1/2\}}$ and $\mathbb{1}_{\{\pi(\mathbf{x})<1/2\}}$ are non-random given $\mathbf{x}$)} \\
&\quad= \mathbb{1}_{\{\pi(\mathbf{x})\geq 1/2\}}\mathbb{P}(y=0 \mid \mathbf{x}=\mathbf{x}) \qquad + \mathbb{1}_{\{\pi(\mathbf{x})<1/2\}}\mathbb{P}(y=1 \mid \mathbf{x}=\mathbf{x}). \\
&\qquad\qquad \text{(Expectation of an indicator random variable)}
\end{aligned}$$

We can distinguish between two cases:

- If $\mathbb{P}(y=0 \mid \mathbf{x}=\mathbf{x}) \geq \mathbb{P}(y=1 \mid \mathbf{x}=\mathbf{x})$ (or $\mathbb{P}(y=0 \mid \mathbf{x}=\mathbf{x}) \geq 1/2$), then any $\pi(\mathbf{x})$ such that $\pi(\mathbf{x})<1/2$ minimizes $\mathbb{E}_{y|x}[L(y,\pi(\mathbf{x})) \mid \mathbf{x}=\mathbf{x}]$.

- If $\mathbb{P}(y = 0 \mid \mathbf{x} = \mathbf{x}) \leq \mathbb{P}(y = 1 \mid \mathbf{x} = \mathbf{x})$ (or $\mathbb{P}(y = 0 \mid \mathbf{x} = \mathbf{x}) \leq 1/2$), then any $\pi(\mathbf{x})$ such that $\pi(\mathbf{x}) \geq 1/2$ minimizes $\mathbb{E}_{y|x}[L(y, \pi(\mathbf{x})) \mid \mathbf{x} = \mathbf{x}]$.

Thus, any π of the form

$$\pi(\mathbf{x}) = \begin{cases} < 1/2, & \text{if } \mathbb{P}(y = 0 \mid \mathbf{x} = \mathbf{x}) \geq 1/2, \\ \geq 1/2, & \text{if } \mathbb{P}(y = 0 \mid \mathbf{x} = \mathbf{x}) < 1/2, \end{cases} \tag{2}$$

minimizes $\mathbb{E}_{xy}\left[L(y, \pi(\mathbf{x}))\right]$ over $\mathcal{H} = \{\pi : \mathcal{X} \to [0, 1]\}$. The posterior distribution $p_{y|x}(1 \mid \mathbf{x})$ is quite naturally of this form, but it is in general not the only π of this kind. As a consequence, the minimizer is not unique.

(Dis-)advantages. The posterior distribution $p_{y|x}$ is the *ground-truth* we seek to find with our (empirical) loss minimization approach. Thus, the corresponding loss function should give an incentive for any learning algorithm to find this ground-truth by minimizing the loss function. This is the case for strictly proper scoring rules like the cross-entropy or log-loss (Bernoulli-loss), but as we have just seen not the case for the probabilistic 0-1-loss. In light of this, it is not a good idea to use the probabilistic 0-1-loss for learning probabilistic classifiers, as the probabilistic classifiers learned might be different from our actual ground-truth posterior distribution. However, one could defend the 0-1 probabilistic loss here as well, since the minimizing probabilistic classifiers in (2) at least have the "correct form" in the sense that the class probabilities are on the right side of 1/2.